Query by image example: the **CANDID** approach

Patrick M. Kelly, Michael Cannon

Computer Research and Applications Group, MS B-265
Los Alamos National Laboratory, Los Alamos, New Mexico

Donald R. Hush

Department of Electrical and Computer Engineering
University of New Mexico, Albuquerque, New Mexico

ABSTRACT

CANDID (Comparison Algorithm for Navigating Digital Image Databases) was developed to enable content-based retrieval of digital imagery from large databases using a query-by-example methodology. A user provides an example image to the system, and images in the database that are similar to that example are retrieved. The development of **CANDID** was inspired by the N-gram approach to document fingerprinting, where a "global signature" is computed for every document in a database and these signatures are compared to one another to determine the similarity between any two documents. **CANDID** computes a global signature for every image in a database, where the signature is derived from various image features such as localized texture, shape, or color information. A distance between probability density functions of feature vectors is then used to compare signatures. In this paper, we present **CANDID** and highlight two results from our current research: subtracting a "background" signature from every signature in a database in an attempt to improve system performance when using inner-product similarity measures, and visualizing the contribution of individual pixels in the matching process. These ideas are applicable to any histogram-based comparison technique.

Keywords: image databases, query by example, image retrieval, probability density functions, histograms

1 Introduction

The retrieval of digital imagery is an active area of research in computer science, promising to provide powerful new tools for database management in the near future. Several products that provide this type of functionality are already available in the marketplace, including Apple's PhotoFlash and IBM's Ultimedia Manager software. These systems allow a user to search image databases using image content as a basis for retrieval, as opposed to employing only textual searches on human-provided keywords or narratives about the database imagery.

In general, image retrieval by content requires algorithms for extracting and comparing color, texture, and/or shape information. Extracted features from the imagery may be associated with entire digital images, or perhaps with specific regions of interest that are identified interactively, semi-automatically, or in a completely automatic manner. The QBIC effort[1,2] is one project that has developed several methods for doing this. As an example, the texture of an image (or of a single object) is represented by a feature vector that can be compared to texture feature vectors from other database images using Euclidean distance, thereby allowing the retrieval of images with "similar" textures.

In contrast to the feature vector approach, color content is typically described using a histogram. A histogram (in the three-dimensional RGB color space) of the colors contained in each image (or in each distinct object) is computed, and an L_1 norm is used to compare these color histograms.[3] Efficient techniques for comparing histograms using quadratic measures of similarity have also been proposed.[4] In this paper, we present a method for retrieving digital imagery by content, where we use probability density functions instead of histograms to describe not only color content, but also localized texture and/or shape content. Our approach was originally motivated by the N-gram method of comparing free-text documents.[5–7]

When using the N-gram method for document comparison, a global signature is computed for each document in a database. This signature represents the content, or topic, of a document in an abstract sense. A signature is typically represented by a histogram of the number of times that each substring of length N occurs in the document, where N is a predetermined value. As an example, for a case-insensitive alphabet of 26 letters, there are 26^3, or 17,576, different tri-grams ("aaa", "aab", "aac", $\cdots$, "zzz"). The signature for each document in this example is therefore a normalized vector of dimension 17,576. A dot-product between N-gram signatures determines the similarity between any two documents. Using this approach for retrieving documents from a database, a user can pose queries such as, "Show me all documents that are similar to this example". A user does not need to identify which specific keywords or phrases are to be searched on; the details of the raw text contain sufficient information for comparing documents.

We have used this idea of comparing global signatures to develop **CANDID** (**C**omparison **A**lgorithm for **N**avigating **D**igital **I**mage **D**atabases), which is analogous to the N-gram approach described above in the sense that we attempt to describe an entire image, or a specific region of interest, with a global signature, and then match signatures with some distance measure to determine image similarity. Every image stored in the database is characterized by a global signature that can represent features such as local textures, shapes, and colors. When a user queries the database to retrieve images that are similar to a given example image, a global signature for that example image is first computed, and this signature is compared to the signatures of all images in the database. All database images are ranked with respect to their similarity to the query image.

In this paper, we present **CANDID**'s methodology and explore some of its potential uses. Section 2 discusses the difference between histogram and probability density function approaches to content comparison. Section 3 talks about different distance / similarity measures that can be used to compare signatures. In Section 4, we present a method for visualizing the similarity between images in a database. Finally, Sections 5 and 6 contain experimental results demonstrating the utility of using our method of query by image example: the **CANDID** approach.

2 Signature Generation

When using a feature-vector approach to describe image content, each component of a feature vector represents a single measurement taken over the entire image (or over an entire region of interest). Histogram approaches, on the other hand, allow us to represent the *distribution of localized features*, instead of restricting us to a single, global measurement. As an illustration of the added utility when using histograms, consider the problem of representing the color content of an image. A feature-vector approach might consist of computing an average red value, an average green value, and an average blue value for the entire image. The result would be a single feature vector of the "average" or "dominant" color characteristics in the image. Using a histogram approach, however, allows us to capture information about the overall distribution of colors in an image.[3] We not only get a feel for the "average" or "dominant" color characteristics of the image, but we also retain information about the relative occurrences of the different color components, such as dark green versus light green. It is therefore generally beneficial to favor histogram approaches over feature-vector based approaches, as long as we consider only the question of information representation and ignore questions concerning efficiency.

Unfortunately, histogram approaches do not scale well with problem dimension. For color representation, where we are concerned with a three-dimensional RGB color space, we can easily divide the space into a discrete number of bins (e.g., 8x8x8 = 512 bins, or 16x16x16 = 4096 bins). As we consider higher-dimensional data, however, the number of bins required for accurate representation grows exponentially. Thus, histogram approaches do not produce viable solutions for problems concerning high-dimensional data. This is one reason that feature-vector approaches to representing information such as texture content are often used.

Using probability density functions to represent the distribution of localized features can circumvent some of the problems associated with higher dimensions. Whereas histograms depend on a discretization of the feature space, probability density functions do not. They can directly represent the distribution of features without specifically designating "bins" in the space. Of course, computing probability density functions is much more expensive than computing histograms, so we are making a sacrifice in terms of computational cost.

CANDID employs probability density functions in an approach that closely resembles the N-gram work for textual data. The general idea is that we first compute several features (local color, texture, and/or shape) at every pixel in the image, and then compute a probability density function that describes the distribution of these features. This probability density function is our content signature for the given image. Of course, probability density function estimation is a large problem in itself; we attempt to estimate the probability density function as a Gaussian mixture. Each Gaussian distribution function is defined by a mean vector $\underline{\mu}_i$ and a covariance matrix Σ_i. A general data clustering routine can provide clusters for which for $\underline{\mu}_i$ and Σ_i can be obtained. We use the k-means clustering algorithm[8,9] followed by an optional cluster merging process.[10] A mean vector and covariance matrix are computed for each of the resultant clusters, and the associated Gaussian distribution function is weighted by the number of elements in the corresponding cluster. Any cluster having a singular covariance matrix is deleted and ignored in subsequent processing. Once a mixture of gaussians has been identified, a signature over a specific N-dimensional feature space for image I can be represented as follows:

$$P_I(\underline{x}) \approx \sum_{i=1}^{K} w_i G_i(\underline{x}) \qquad ; \qquad G_i(\underline{x}) = (2\pi)^{-\frac{N}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x}-\underline{\mu}_i)\right] \tag{1}$$

We differentiate between two different signatures, $P_{I_1}(\underline{x})$ and $P_{I_2}(\underline{x})$, with the following notation:

$$P_{I_1}(\underline{x}) = \sum_{i=1}^{K_1} w_i G_i(\underline{x}) \qquad P_{I_2}(\underline{x}) = \sum_{j=1}^{K_2} v_j F_j(\underline{x}) \tag{2}$$

3 Signature Comparison

There are several comparison functions that we can use to compare two signatures. If we are comparing histograms, for example, we could easily use an L_1 distance measure. Consider using an L_1 distance measure to compare continuous probability density functions represented by Gaussian mixtures. We will use our signature representation from Equation (2):

$$dist_{L_1}(I_1, I_2) = \int_{\Re} |\, P_{I_1}(\underline{x}) - P_{I_2}(\underline{x}) \,|\, d\underline{x} \tag{3}$$

$$= \int_{\Re} \left| \sum_{i=1}^{K_1} w_i G_i(\underline{x}) - \sum_{j=1}^{K_2} v_j F_j(\underline{x}) \right| d\underline{x} \tag{4}$$

A closed-form solution for calculating this distance measure is not obtainable due to the discontinuity caused by the absolute value function. We can, however, obtain a closed-form solution if we use an L_2 distance measure:

$$dist_{L_2}\,(I_1, I_2) \;=\; \left[\int_{\Re} \left(P_{I_1}(\underline{x}) - P_{I_2}(\underline{x}) \right)^2 d\underline{x} \right]^{\frac{1}{2}} \tag{5}$$

We have previously discussed a normalization of this distance measure,[11] which guarantees that distances will always be between 0 and 1. We refer to this "normalized" distance as $ndist_{L_2}(I_1, I_2)$.

We can also use an inner product to compare histograms, which is applicable to comparing continuous probability density functions as well. This inner product is referred to as a similarity measure, where a result of 0 indicates *no similarity*:

$$sim\,(I_1, I_2) \;=\; \int_{\Re} P_{I_1}(\underline{x}) P_{I_2}(\underline{x}) d\underline{x} \tag{6}$$

If we choose to normalize this similarity measure, then we can interpret the result as being the cosine of the angle between the two functions:

$$nsim\,(I_1, I_2) \;=\; \frac{\int_{\Re} P_{I_1}(\underline{x}) P_{I_2}(\underline{x}) d\underline{x}}{\left[\int_{\Re} P_{I_1}^2(\underline{x}) d\underline{x} \int_{\Re} P_{I_2}^2(\underline{x}) d\underline{x} \right]^{\frac{1}{2}}} \tag{7}$$

The four previous functions can be used to compare signatures that are represented by Gaussian mixtures, and closed-form solutions are available.[11,12]

As suggested by recent N-gram research, we have considered subtracting a dominant "background" from every signature in our database prior to signature comparison.[7] If we are using a true distance function to compare signatures, then this subtraction does nothing to affect the comparisons. If, however, we are considering a similarity measure such as $nsim\,(I_1, I_2)$, then subtracting a dominant background has a dramatic effect on our comparisons. Subtracting the "average" of all signatures in the database from every signature compensates for the bias caused by those features that are common to all images; this "background" may be unimportant when comparisons are being made, and it should not overpower the "details" that we want to exploit.

Given our method for representing signatures, namely as continuous probability density functions, it is impractical to mathematically subtract an average probability density function from all database signatures and store that result for later use. It is much easier to compute a background signature $P_{BG}(\underline{x})$ (by whatever process is reasonable for the problem at hand) and incorporate this background directly into our similarity measure. As an example, if we are interested in subtracting a background signature from each database signature prior to comparisons with our normalized similarity measurement $nsim\,(I_1, I_2)$, we can manipulate Equation (7) accordingly:

$$\begin{aligned}
nsimbg\,(I_1, I_2) \;&=\; \frac{\int_{\Re} \left(P_{I_1}(\underline{x}) - P_{BG}(\underline{x})\right) \left(P_{I_2}(\underline{x}) - P_{BG}(\underline{x})\right) d\underline{x}}{\left[\int_{\Re} \left(P_{I_1}(\underline{x}) - P_{BG}(\underline{x})\right)^2 (\underline{x}) d\underline{x} \int_{\Re} \left(P_{I_2}(\underline{x}) - P_{BG}(\underline{x})\right)^2 (\underline{x}) d\underline{x} \right]^{\frac{1}{2}}} \\
&=\; \left[\int_{\Re} P_{I_1}(\underline{x}) P_{I_2}(\underline{x}) d\underline{x} + \int_{\Re} P_{BG}^2(\underline{x}) d\underline{x} - \int_{\Re} P_{I_1}(\underline{x}) P_{BG}(\underline{x}) d\underline{x} - \int_{\Re} P_{I_2}(\underline{x}) P_{BG}(\underline{x}) d\underline{x} \right] \cdot \\
&\quad \left[\int_{\Re} P_{I_1}^2(\underline{x}) d\underline{x} - 2 \int_{\Re} P_{I_1}(\underline{x}) P_{BG}(\underline{x}) d\underline{x} + \int_{\Re} P_{BG}^2(\underline{x}) d\underline{x} \right]^{-\frac{1}{2}} \cdot \\
&\quad \left[\int_{\Re} P_{I_2}^2(\underline{x}) d\underline{x} - 2 \int_{\Re} P_{I_2}(\underline{x}) P_{BG}(\underline{x}) d\underline{x} + \int_{\Re} P_{BG}^2(\underline{x}) d\underline{x} \right]^{-\frac{1}{2}}
\end{aligned} \tag{8}$$

4 Visualization of Results

One of the problems with query-by-example information retrieval systems is that the result of a query is simply a group of items that are hopefully interesting to the user (in our case, a group of images that are "similar" to the query image). Some additional information, such as similarity scores produced by the comparison process, might also be returned to allow a user to gauge the "correctness" of the result. It is reasonable for a user to pose questions such as, "Why do these two images look similar?" or "What specific parts of these images are contributing to the similarity?". Techniques where pixels are assigned to bins (histogram approaches) or to clusters (probability density function approaches) retain information about pixel-membership relationships that can be used to produce visualizations that are helpful in answering these questions. We can use our probability density function signatures to display results in such a way as to indicate exactly where the matching occurred. If we re-arrange our equation for $sim(I_1, I_2)$, we can isolate K_1 different terms that contribute to the overall similarity score:

$$\begin{aligned} sim(I_1, I_2) &= \int_{\Re} P_{I_1}(\underline{x}) P_{I_2}(\underline{x}) d\underline{x} \\ &= \int_{\Re} \left[\sum_{i=1}^{K_1} w_i G_i(\underline{x}) \cdot \sum_{j=1}^{K_2} v_j F_j(\underline{x}) \right] d\underline{x} \\ &= \sum_{i=1}^{K_1} w_i \int_{\Re} \left[G_i(\underline{x}) \sum_{j=1}^{K_2} v_j F_j(\underline{x}) \right] d\underline{x} \\ &= \sum_{i=1}^{K_1} contrib_i \qquad (9) \end{aligned}$$

Each of the K_1 gaussians in $P_{I_1}(\underline{x})$ makes some contribution to $sim(I_1, I_2)$. Furthermore, every pixel in image I_1 was assigned to exactly one cluster (i.e. Gaussian distribution) during signature generation. We can therefore build a new image where every pixel is highlighted in a manner consistent with the contribution made to $sim(I_1, I_2)$ by the cluster to which this pixel was assigned. We can use this method for visualizing exactly what parts of the two images, I_1 and I_2, contributed to the match. The normalized similarity measure, $nsim(I_1, I_2)$, can be decomposed in a similar manner.

5 Experimental Results: Landsat TM Data

Remotely-sensed data can be used to locate underground oil reserves, monitor pollution from large factories, and track the disappearance of our world's rain forests. A database containing imagery collected by airborne sensors will prove much more valuable if scientists can access the data by searching on different attributes of image content instead of only being able to retrieve data by searching on associated textual metadata information. The ability to automatically locate areas having similar ground cover will enable scientists to search through terabyte-sized image databases in order to study environmental problems. As an example, if a coniferous forest in Oregon is rapidly disappearing for no apparent reason, then other areas around the world having similar vegetation can be retrieved to see if they are experiencing the same problem. Scientists would then know if this was a global phenomenon or if local conditions were to blame.

We have applied **CANDID** to the problem of retrieving multispectral satellite data (Landsat TM data) from a database. This enables queries such as, "Show me all images of areas with landcover similar to this example." As an experiment, we created a database containing 100 512 × 512, 6-banded images (the thermal infrared band in each image was ignored). The sample images used to populate our database were acquired from four different

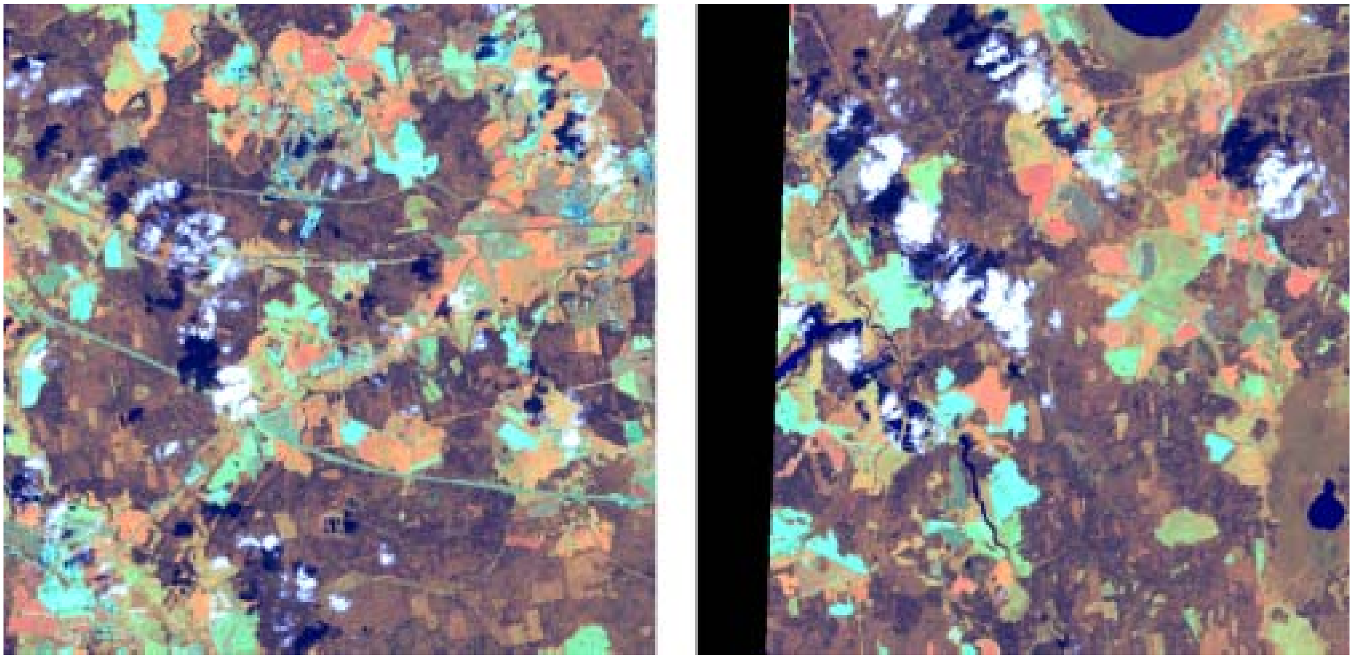

Figure 1: Retrieval of Landsat TM data. The query image on the left was selected from the Moscow scene. Using signatures represented by 20 gaussians in the 6-dimensional feature space, **CANDID** identified the image on the right as being the best match from the other database images (with a match score of 0.89).

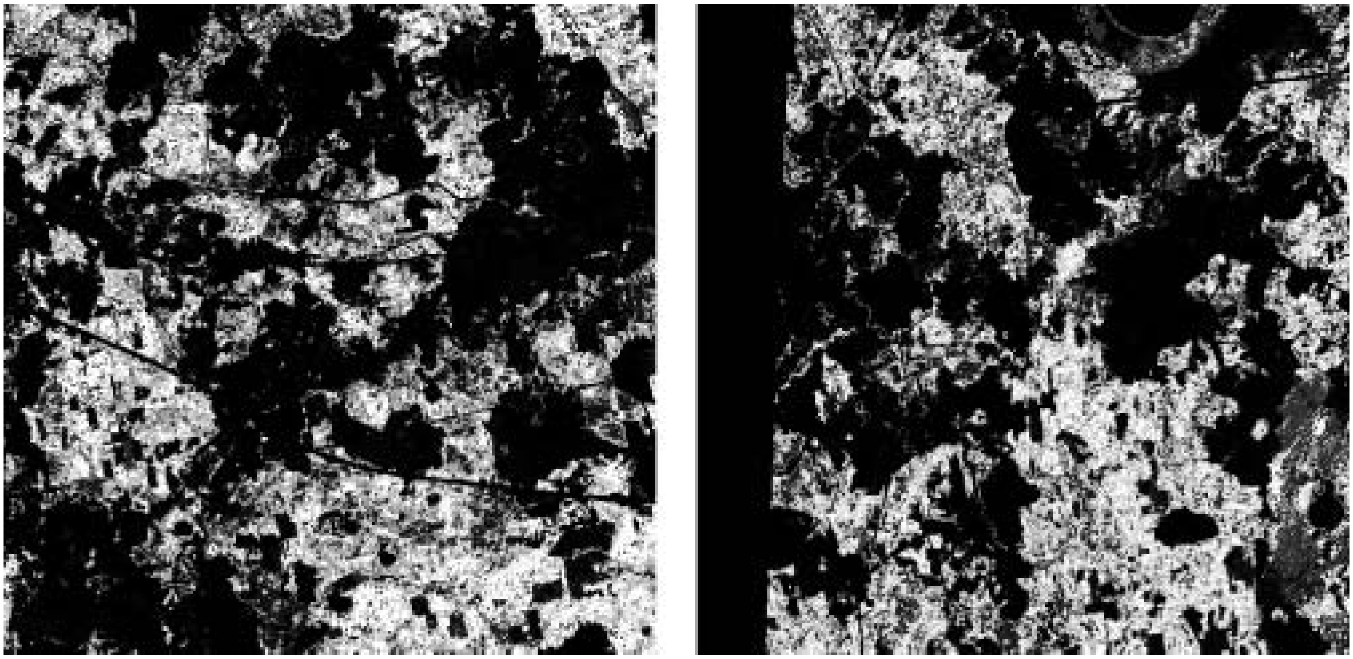

Figure 2: Pixels Contributing to Match. Pixels in these images are shaded according to their overall contribution to the high similarity score between the two images in the previous figure. White pixels contributed the most.

geographic locations, each having its own characteristic landscape (see Table 1). The Moscow area, for example, contains many diverse landcover types in every 512 × 512 subimage that was extracted. These landcover types include coniferous forest, deciduous forest, and agriculture. The Moscow images look nothing like the images around the other three geographic locations. Similarly, the Cairo landscape is unique and dissimilar to the Moscow, Albuquerque, and Los Alamos areas.

LOCATION	DOMINANT LANDSCAPE COVER
Moscow (Russia)	Coniferous Forest, Deciduous Forest, Agriculture, ...
Cairo (Egypt)	Agriculture, Dense Urban, ...
Albuquerque (USA)	Desert, Coniferous Forest, ...
Los Alamos (USA)	Desert, Coniferous Forest, ...

Table 1: Selected Geographic Locations

We calculated global spectral signatures for each database image by clustering the 6-dimensional pixel vectors into 20 clusters. We then used **CANDID** to query our database using an example image from the Moscow scene. We used our normalized similarity measure, $nsim(I_1, I_2)$, and sorted the similarity scores between the query image and all 100 test images in the database, which we then plotted (see Figure 3). All subimages from the Moscow area were retrieved before subimages around Cairo, Albuquerque, and Los Alamos. Furthermore, all Moscow images in the database yielded similarity scores between 0.5 and 1.0, whereas the other images produced similarity scores below 0.01. The point is that using a query image from the Moscow scene, all other images from the Moscow scene (which are all of similar landscape) are retrieved from the database first, after which the other images (from the Cairo, Albuquerque, and Los Alamos scenes) are retrieved with negligible match scores. Figure 1 shows the Moscow query image along with the best match from the database. Using the visualization method described in Section 4, pixels contributing to the match were highlighted, and the resultant images are displayed in Figure 2, with white pixels contributing the most.

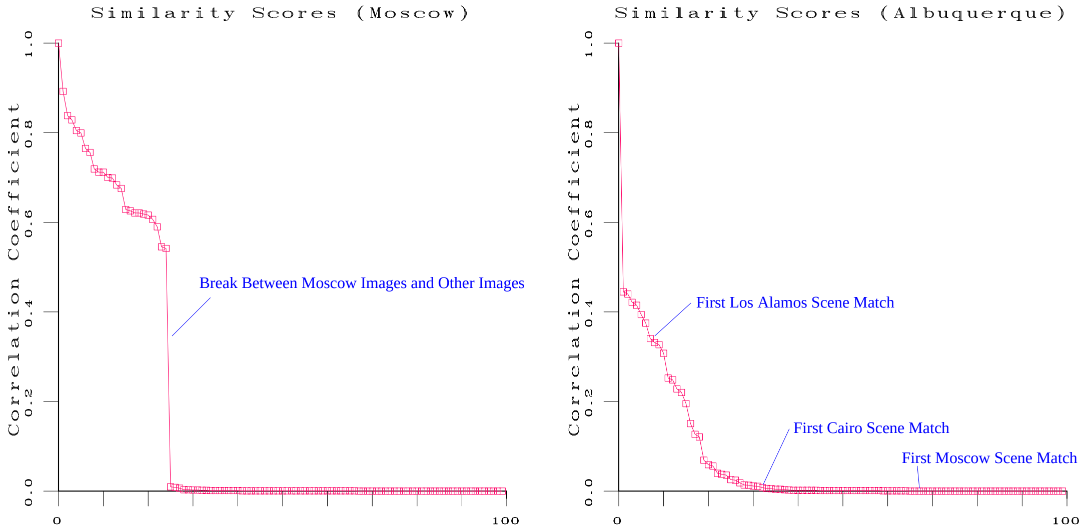

Figure 3: Sorted similarity scores using example images from the Moscow scene and from the Albuquerque scene. The Moscow example produced match scores greater than 0.5 when compared to all other Moscow images, while producing match scores under 0.01 when compared to all other images. The Albuquerque example, on the other hand, did not produce any match scores greater than 0.5 (except when compared with itself).

Unlike Moscow and Cairo, the Albuquerque and Los Alamos images vary quite a bit; 512×512 subimages that were adjacent to one another in the original data set do not necessarily have a lot in common. This is because the New Mexico landscape contains a combination of desert areas, mountainous areas (dominated by coniferous forest), and transitional areas. Again, we used **CANDID** to search the database for images similar to a query image from the Albuquerque scene. The sorted similarity scores are plotted in Figure 3. This plot shows that images from both the Albuquerque and Los Alamos scenes are retrieved first, but the corresponding similarity scores are typically low (less than 0.5). This reflects the fact that images in both the Albuquerque and Los Alamos scenes contain some geographically similar data, but they are not as homogeneous as, say, the Moscow scene. Images from the Moscow and Cairo scenes, which are not at all similar to the query image, are retrieved with negligible match scores. Results from these experiments were quite promising, indicating that further study into the application of **CANDID** to image retrieval for remote-sensing problems is warranted.

6 Experimental Results: Pulmonary CT Data

We have used the concept of global signature matching to retrieve medical imagery based on image content. Pulmonary CT scans reveal the gross pathology indicative of diseased lung tissue resulting from a variety of disorders such as lymphangioleiomyomatosis (LAM), idiopathic pulmonary fibrosis (IPF), scleroderma, emphysema, asthma, and vasculitis. Since CT data is acquired digitally, it can be easily stored in a computer database. It would be a natural extension of this process to search a database to retrieve images that exhibit the same pathology as the current study. These images would provide the radiologist with immediate access to past cases where similar problems were encountered, thereby aiding with the current patient's diagnosis and treatment.

Diagnosis	Number of Patients	Total Number of Images
LAM	11	46
Scleroderma	1	20
IPF	6	12
Emphysema	2	10
Normal	3	6
Vasculitis	1	46
Asthma	10	80
TOTALS	34	220

Table 2: Contents of CT Image Database

We applied **CANDID** to this problem of retrieving pulmonary CT imagery from a database containing a total of 220 lung images taken from pulmonary CT studies of 34 different patients (see Table 2). Each image was 512×512 pixels in size, consisting of 12-bit grayscale data. For this application, we are primarily interested in retrieving images containing similar textures. We have previously demonstrated that four Laws texture energy measures[13,14] can be used to discriminate between images of lungs that are affected by by different diseases.[11,12] A drawback to using Laws texture energy measures is the amount of time it takes to generate features that can be submitted to a clustering algorithm for signature generation. We have recently been experimenting with other features that contain local texture information. One of these feature sets, which we simply call "local statistics", is very fast to compute as compared to the Laws texture energy measures.

Before we computed any statistical features, we first isolated those areas of each CT image that were representative of lung tissue,[15] and we worked only within these areas. The local statistics data set consisted of three features computed on the grayscale values surrounding each lung pixel: standard deviation (σ), skewness, and kurtosis. The pixel intensity values within a circular region of diameter 9 around each pixel were extracted (surrounding pixels falling outside of the lung were ignored). Three statistical features were computed at each pixel from these surrounding grayscale values, using definitions consistent with modules available in the Khoros platform.[16] In the equations below, μ denotes the mean value of the pixels in the circular region.

$$\sigma = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2\right]^{\frac{1}{2}} \qquad skewness = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^3 \qquad kurtosis = \frac{1}{N\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^4 - 3 \quad (10)$$

Each lung pixel is now represented by a feature vector containing three localized statistical measurements. We cluster these vectors into 20 clusters to build signatures for each image in our three-dimensional feature space.

We also computed a baseline signature for the entire data set to be used as a "background" to be subtracted from all individual signatures. This was done by computing the three-dimensional statistical feature vectors for all lung pixels in every CT image in our database, and then randomly extracting ten percent of these feature vectors to serve as a background data set. We clustered this background data set to 20 clusters to produce our background data signature.

In order to compare the performance of the six different distance/similarity measures mentioned in Section 3, we devised an experiment based upon our database of 220 pulmonary CT images. We assume that any image that comes from the same CT study as the query image should be ranked as *very* similar to the query image. The images displayed in Figure 4 come from a single CT study of a patient with Lymphangioleiomyomatosis, a disease characterized by the formation of empty cysts in the lung. We used the image in the upper-left corner to query our CT database, after which we ranked all database images by their relative similarity to that query image. We used six distance/similarity measures, and we compare their effectiveness by noting how well each was able to retrieve the images in Figure 4. The results of the experiment are shown in Table 3.

We note that the two distance measures, both normalized and non-normalized, as well as the normalized similarity with background subtraction, performed with great accuracy. That is, the best matches produced by these measures were from the same CT study as the query image.

	Rankings Produced By					
	$dist_{L_2}$	$ndist_{L_2}$	sim	$nsim$	$simbg$	$nsimbg$
Slice 1 (Query Image)	1	1	1	1	1	1
Slice 2	18	46	151	93	17	16
Slice 3	5	4	111	6	33	5
Slice 4	4	6	127	14	16	4
Slice 5	3	3	92	3	30	3
Slice 6	2	2	94	2	26	2

Table 3: The six CT slices in the table were retrieved from our pulmonary CT database using six different distance/similarity measures. Each column shows the retrieval rankings of the individual images resulting from a search based on a given measure. Those measures that retrieved images with high rankings (1 being the highest), are judged to be superior to those that ranked the images relatively lower.

7 Conclusions

A general approach to the problem of comparing images with respect to localized color, texture, or shape information has been proposed, and promising results for two applications have been presented. The benefit to using this approach is that information is captured in a manner similar to histogram approaches, but it is not limited by a need for identifying discrete bins. At present, the **CANDID** approach is not viable for large problems as it does not scale well. Probability density functions are convenient in terms of information representation and

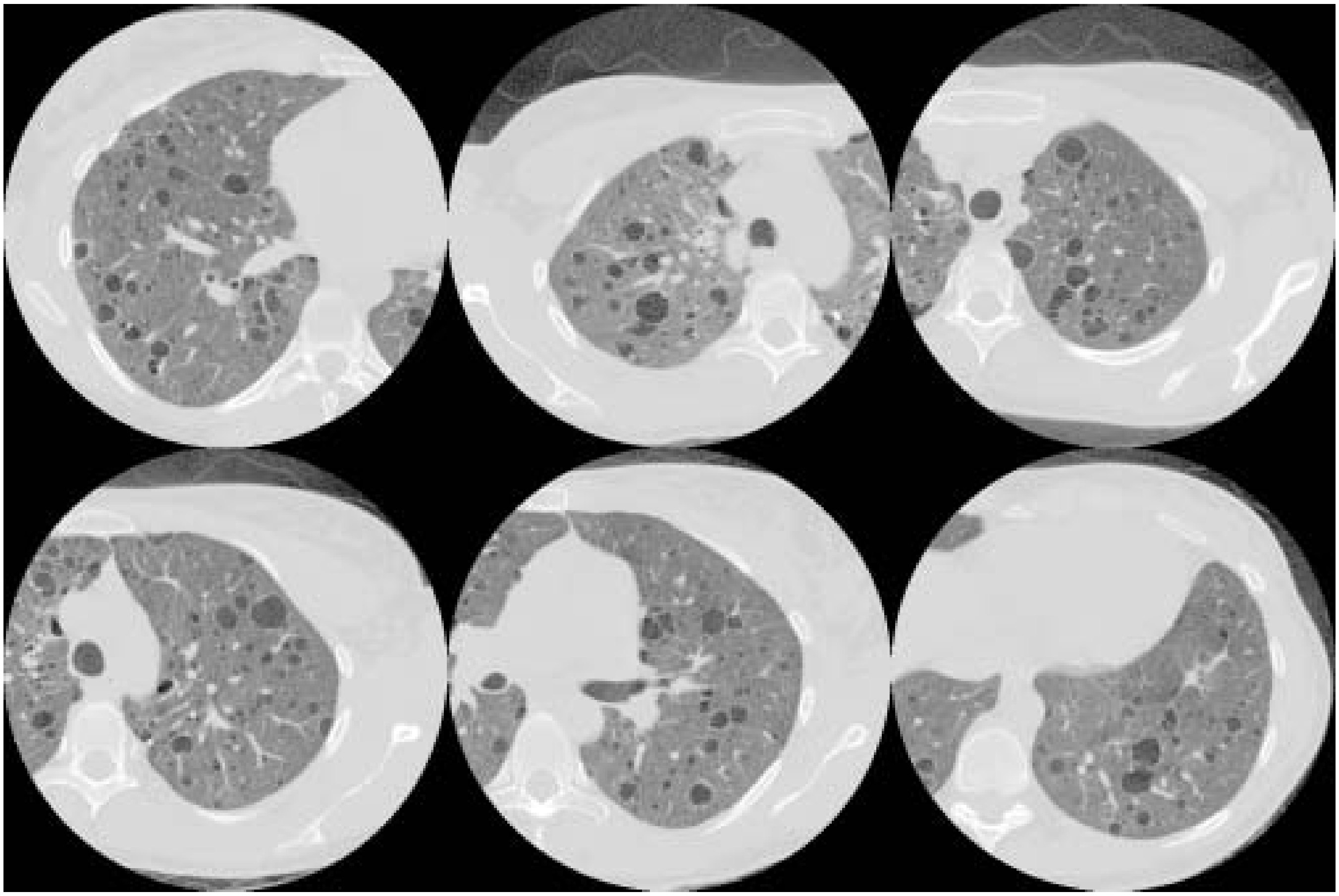

Figure 4: Lymphangioleiomyomatosis (LAM). These six images come from a single CT study done for a patient with LAM disease, which is characterized by the formation of empty cysts in the lung. When one of these images is used to query our database, we would "expect" that the other images from this same study should be retrieved early.

comparison, but these methods are slower than other techniques. A need for efficient indexing methods for **CANDID** is apparent.

Some of the recent results from this project, such as subtracting background signatures and visualizing individual pixel contributions, are also applicable to histogram-based comparison techniques. Subtracting a background signature from every signature in a database, as proposed in the literature concerning the retrieval of free-text documents, improves system performance when inner-product similarity measures are considered. Visualizing the contribution of individual pixels in the matching process is possible by considering the cluster assignments (or bin assignments) of each pixel.

8 Acknowledgements

Dr. John Newell and Dr. David Lynch at National Jewish Center for Immunology and Respiratory Medicine provided the CT data used to perform the medical image comparison work. Calvin Hamilton provided programs to

extract local statistical measures from grayscale pixel values. This work was performed under a U.S. Government contract (W-7405-ENG-36) by Los Alamos National Laboratory, which is operated by the University of California for the U.S. Department of Energy.

9 REFERENCES

[1] W. Niblack, R. Barder, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Yaubin. The QBIC project: Querying images by content using color, texture, and shape. In *SPIE Vol. 1908 Storage and Retrieval for Image and Video Databases*, pages 173–181, 1993.

[2] C. Faloutsos, M. Flickner, W. Niblack, D. Petkovic, W. Equitz, and R. Barber. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, pages 231–262, 1994.

[3] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[4] H.S. Sawhney and J.L. Hafner. Efficient color histogram indexing. In *Proceedings of the 1994 IEEE International Conference on Image Processing*, volume 2, pages 66–70, 1994.

[5] R.E. Kimbrell. Searching for text? send an n-gram! *BYTE*, pages 297–312, May 1988.

[6] T.R. Thomas. Document retrieval from a large dataset of free-text descriptions of physician-patient encounters via n-gram analysis. Technical Report LA-UR-93-0020, Los Alamos National Laboratory, Los Alamos, NM, 1993.

[7] M. Damashek. Gauging similarity via n-grams: Text sorting, categorization, and retrieval in any language. *submitted to Science*. In review.

[8] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, NY, 1973.

[9] J.T. Tou and R.C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley, Reading, MA, 1974.

[10] P.M. Kelly, D.R. Hush, and J.M. White. An adaptive algorithm for modifying hyperellipsoidal decision surfaces. *Journal of Artificial Neural Networks*, 1(4). In Press.

[11] P.M. Kelly and T.M. Cannon. CANDID: Comparison algorithm for navigating digital image databases. In *Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management*, pages 252–258, 1994.

[12] P.M. Kelly and T.M. Cannon. Experience with CANDID: Comparison algorithm for navigating digital image databases. In *SPIE Vol. 2368 Proceedings of the 23rd AIPR Workshop on Image and Information Systems: Applications and Opportunities*, 1994. To appear.

[13] K. Laws. *Textured Image Segmentation*. Ph.D. dissertation, Univ. of Southern Calif., January 1980.

[14] K. Laws. Rapid texture identification. In *SPIE Vol. 238 Image Processing for Missile Guidance*, pages 376–380, 1980.

[15] J. Everhart, M. Cannon, J. Newell, and D. Lynch. Image segmentation applied to ct examination of lymphangioleiomyomatosis (lam). In *SPIE Vol. 2167 Medical Imaging 1994: Image Processing*, pages 87–95, 1994.

[16] J. Rasure and C. Williams. An integrated visual language and software development environment. *Journal of Visual Languages and Computing*, 2:217–246, 1991.